기계 독해 데이터셋을 이용한 Dense Passage Retrieval

서민택 나승훈
전북대학교

신동욱 김선훈 강인호
네이버

KCC22

Introduction

- 개방형 도메인 작업을 위해서는 검색이 선행 되어야한다.
- 최근 희소 벡터를 이용한 방법에서 밀집 벡터를 이용한 방법으로 전환되고 있음.
- 한국어 적용을 위해 모델을 학습하기엔 한국어 검색 데이터가 적음.
- 많이 공개 되어 있는 기계 독해 데이터셋을 통해서 밀집 벡터 검색 모델 학습이 충분히 가능함을 보여줌.

희소 검색 vs 밀집 검색

- 위와 같이 의미적으로 연관이 있더라도 희소적 방법은 형태가 다르다면 검색이 어려움
- 따라서 밀집 벡터를 이용한 표현을 통하여 의미적은 내용을 이용하도록 모델을 학습

ICT Style vs DPR Style

Passage

1989년 3월 12일 서울지방검찰청 공안부는 임종석의 사전구속영장을 발부받았다. 같은 해 6월 30일 평양축전에 임수경을 대표로 파견하여 국가보안법위반 혐의가 추가되었다...

- **ICT Style Qurey**
  1989년 3월 12일 서울지방 검찰청 공안부는 임종석의 사전 구속 영장을 발부받았다. (내부 문장 그대로 활용)

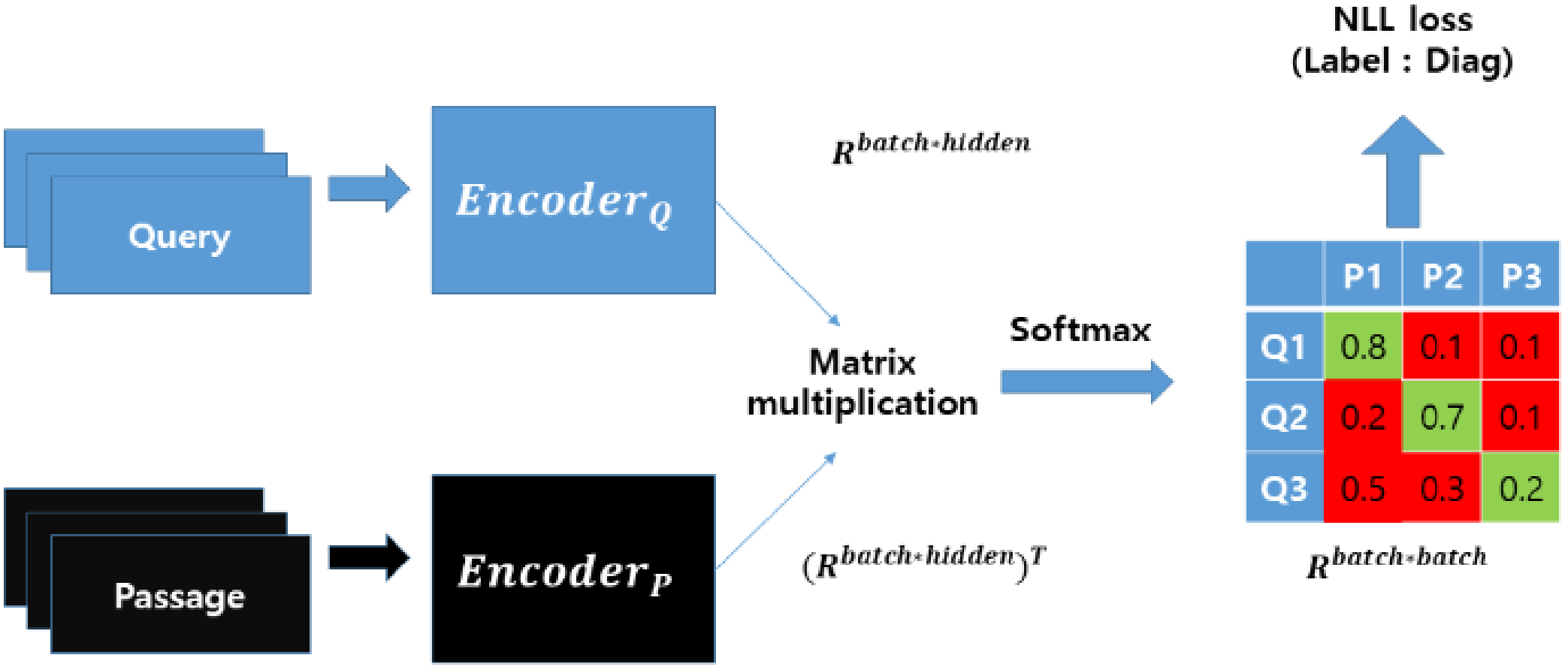
- **DPR Style Qurey**
  임수경을 평양축전 대표로 파견한 사람은 누구야? (일반적 질의)

기계 독해 데이터 예시

- 질문 : 바그너는 교향곡 작곡을 어디까지 쓴 뒤에 중단 했는가?
- 정답 : 교향곡
- 문서 : 1839년 바그너는 괴테의 파우스트를 읽고.... (Wiki 단락)

- 정답을 제거하면 질문-문서 쌍으로 이루어진 DPR Style의 데이터가 구성됨
- Korquad 1.0 데이터를 이용

Dense Passage Retrieval



성능 증가를 위한 시도

- **요약 모델 Zero-Shot 데이터 증강**

- >단락을 파싱 후 요약문 생성 모델을 이용해 Query를 생성하여 데이터의 크기를 늘린다

- **Batch Size 증가**

- > In-Batch Negative의 방식에 의해 Batch 크기가 증가는 곧 Negative Sample의 증가함.

성능 비교

Recall@k	R@1	R@3	R@5	R@10	R@20	R@50	R@100
Realm	45.07	63.85	70.89	76.53	81.22	88.73	92.02
DPR[D]	13.14	22.06	28.63	39.43	45.07	57.74	64.78
DPR[S]	34.71	56.33	66.17	73.23	78.87	85.91	91.54
DPR[A]	37.55	64.33	71.36	81.69	88.26	91.54	93.42
Zeroshot+	32.86	55.86	70.04	76.52	84.97	90.61	92.01
BatchSize+	37.08	63.84	72.72	82.15	90.14	95.30	97.18

D : 인코더 분리　　S : 인코더 가중치 공유　　A : 400K 데이터 추가(AIHUB)

결론

- 기계 독해 데이터를 이용해도 준수한 성능의 검색기를 획득가능
- 한국어는 인코더를 공유하는 방식이 좀 더 좋은 성능을 보여줌
- 단 R@1의 성능이 ICT 방법보다 낮음.
- 하지만 그 이상의 성능이 높기 때문에 사실 검증 같은 여러 문서가 필요한 작업에서 좀더 좋은 성능 기대
- 엄밀하지 않은 ZeroShot 데이터 추가는 성능의 하락을 가져옴
- Batch size의 증가는 Negative Sample의 증가를 가져와 성능 향상